Ответы на вопросы

Вопрос 1

Какие методы увеличения данных Вы знаете?

Я полагаю, в вопросе имеются ввиду методы увеличения количества данных для обучения. Если в пространтве обьектов допускаются преобразования, сохраняющие инвариантные признаки, то можно воспользоваться ими для увеличения количества данных. Например, для картинок, в случае задачи распознавания, можно добавлять небольшие шумы, использовать повороты, отражение и переносы изображения. Также можно уменьшить влияние несущественных признаков, деформируя, масштабируя, изменяя яркость, цвет картинок и т.д.
Если мы имеем дело с текстом, то можно заменять слова на синонимы, вставлять в предложения слова, не рушащие правильную грамматическую структуру текста и т.д.

Вопрос 2

Что такое бутстреп?

Бутстрэп - это статистический метод, который позволяет оценивать многие статистики сложных распределений.
Метод бутстрэпа заключается в следующем. Пусть имеется выборка X размера N. Равномерно возьмем из выборки N объектов с возвращением. Обозначим новую выборку через X_1. Повторяя процедуру M раз, сгенерируем M подвыборок X. Теперь мы имеем достаточно большое число выборок и можем оценивать различные статистики исходного распределения.
Один из самых первых и самых простых видов ансамблей основан на бутстрепе.
Bagging (от Bootstrap aggregation):
Пусть имеется обучающая выборка X. С помощью бутстрэпа сгенерируем из неё выборки $X_1, X_2, \ldots, X_M$. Теперь на каждой выборке обучим свой классификатор $a_i\left(x\right)$. Итоговый классификатор будет усреднять ответы всех этих алгоритмов.
Средняя квадратическая ошибка при использовании беггинга уменьшается в M раз.
Беггинг позволяет снизить дисперсию (variance) обучаемого классификатора, уменьшая величину, на сколько ошибка будет отличаться, если обучать модель на разных наборах данных, или другими словами, предотвращает переобучение. Эффективность бэггинга достигается благодаря тому, что базовые алгоритмы, обученные по различным подвыборкам, получаются достаточно различными, и их ошибки взаимно компенсируются при голосовании, а также за счёт того, что объекты-выбросы могут не попадать в некоторые обучающие подвыборки.